

# Predicting Climate Variability over the Indian Region Using Data Mining Strategies

M. Naresh Kumar<sup>a,\*\*</sup>

<sup>a</sup>*Software and Database Systems Group, National Remote Sensing Center (ISRO),  
Hyderabad, Telangana, 500 037, India*

---

## Abstract

In this paper an approach based on expectation maximization (EM) clustering to find the climate regions and a support vector machine to build a predictive model for each of these regions is proposed. To minimize the biases in the estimations a ten cross fold validation is adopted both for obtaining clusters and building the predictive models. The EM clustering could identify all the zones as per the Koppen classification over Indian region. The proposed strategy when employed for predicting temperature has resulted in an RMSE of 1.19 in the Montane climate region and 0.89 in the Humid Sub Tropical region as compared to 2.9 and 0.95 respectively predicted using k-means and linear regression method.

*Keywords:* support vector machine, expectation maximization, k-means, regression, climate regions, climate change, Koppen classification

---

## 1. Introduction

Regionalization techniques are found to be effective in improving the prediction accuracies of the climate models. Building regional models and predicting the climate variability require processing and extraction of information from large volumes of high dimensional data sets. Data mining methods such as k-means (KM) clustering and statistical methods such as linear regression (LR) are popular techniques commonly employed for grouping the

---

\*Principal Corresponding Author

\*\*Tel.: +91 40 2388 4388; Fax.: +91 40 2388 4437

Email address: [nareshkumar\\_m@nrsc.gov.in](mailto:nareshkumar_m@nrsc.gov.in) (M. Naresh Kumar)

data into regions of similar climate and build a model to predict the climate variables for subsequent years. The k-means method requires specifying initial k clusters centers which is generally not known a priori. Also, the procedure is sensitive to the selection of the initial cluster centers. Moreover, a linear regression model may not capture the non-linear relationships among the climate variables.

The EM finds clusters by finding a appropriate fit for the given data set with a mixture of Gaussians. Each of the Gaussians is associated with a mean and a covariance matrix. The prior probability for each Gaussian is computed as a total fraction of points in the cluster defined by that Gaussian. Based on the iterative approach in updating values for means and variances the optimal solution is reached.

In this paper an approach based on expectation maximization (EM) clustering to find the climate regions and a support vector machine to build a predictive model for each of these regions is proposed. To minimize the biases in the estimations a ten cross fold validation is adopted both for obtaining clusters and building the predictive models.

The following are the main objectives of the present work

1. Understand the process of climate change over Indian region through development of information extraction techniques that can effectively predict the climate variability
2. Develop a methodology for processing the long term gridded climate data and obtaining climate regions using expectation maximization clustering
3. Prepare the maps of the climate regions identified by expectation maximization clustering and compare it with standard climate zones as per Kppen classification over Indian region
4. Evolve a procedure to subset the long term climate dataset into regional data sets
5. Develop methods to extract the train data set for building the support vector regression classifier based on the number of years to predict
6. Obtain the validation data set for each of the grid locations and compute the root mean squared error.
7. Compare the performance of the proposed methodology with k-means and linear regression procedure

This paper is organized as follows. In Section 2 the proposed methodology

of predicting climate variables is presented. The experiments and results are discussed in Section 3. Conclusions and discussion are deferred to Section 4.

## 2. Methodology

In the proposed methodology the climate dataset is first regionalized by applying Expectation maximization clustering using the long term averages of the climate variables. Further, a predictive model is developed using support vector machine SVM regression kernel. A ten cross fold validation is employed to obtain a robust estimates of the root mean square error (RMSE).

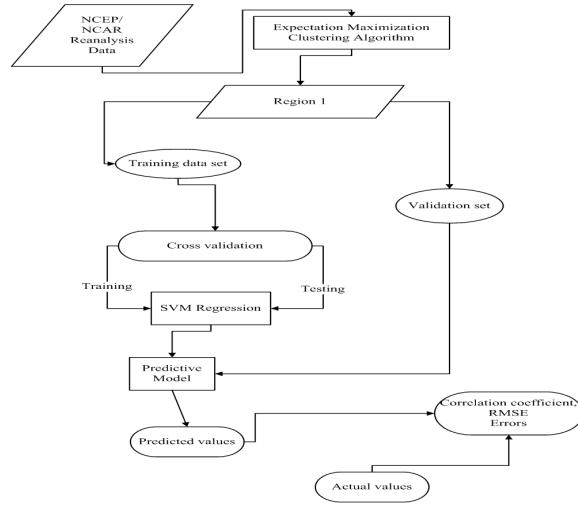


Figure 1: A flow chart depicting the procedure for building a predictive model

The procedure employed in developing a predictive model is shown in Figure 1.

The algorithm 1 describes the steps implemented in the present paper for obtaining a model for predicting climate variables.

## 3. Experiments and Results

NCEP/NCAR reanalysis data for 65 years from 1948 to 2012 having the climate variables Atmospheric Pressure, Relative Humidity, Precipitable Water, Zonal Wind, Meridional Wind, Precipitation, Air Temperature is used in the analysis

---

**Algorithm 1** Procedure for Predicting Climate Variables

---

**Require:** 1. Climate data set  
2. Clustering method  
3. Number of years to predict ( $p$ )  
4. Variable to be predicted

**Ensure:** 1. Correlation coefficient  
2. Root mean squared error

**Algorithm**

1. Extract long term mean of climate variables for each  $2.5^\circ \times 2.5^\circ$  grid over Indian region
  2. Apply clustering method to obtain regions  $R_1, R_2, \dots, R_n$
  3. Build the Model for the variable to be predicted
    - (a) For each region in  $R_1, R_2, \dots, R_n$ 
      - i. obtain mean of the climate variables for all the grid points in the cluster for  $j-p$  years where  $j$  denotes total number of years and  $p$  denotes number of years for which prediction is required
      - ii. build a support vector machine regression model using a ten cross fold validation procedure
  4. Test the model built in Step 3
    - (a) For each cluster in  $R_1, R_2, \dots, R_n$ 
      - i. For each grid point in the cluster
        - A. apply the corresponding model to predict precipitation and temperature for years  $1, \dots, p$
        - B. compute the  $RMSE$  using the predicted values and the actual values of the climate variables
  5. RETURN  $RMSE$ .
  6. END.
-

Table 1: EM Cluster Centriods of different Climate Zones

Climate Variable	Montanenew	Semi Arid	Tropical Wet and Dry	Arid	Montane	Tropical Wet	Humid Sub tropical
Air Temperature	12.44	27.04	25.79	25.81	-2.54	26.83	24.8
Precipitable water	18.86	41.32	29.19	22.02	6.31	38.01	37.61
Precipitation	4.8	3.1	2.57	0.67	2.36	3.03	6.4
Relative Humidity	81.96	76.73	53.05	35.19	78.81	75.51	74.5
Sea Level Pressure	1011.07	1008.87	1008.08	1007.8	1015.22	1009.76	1009.43
Zonal winds	0.69	0.94	0.57	1.1	2.99	2.67	0.69
Meridional winds	1.02	1.63	-0.36	0.88	1.82	-1.01	0.54

The application of the EM clustering on the dataset has resulted in 7 climate regions. As per Koppen Classification only there are six regions. The present algorithm 1 has brought out a new region consisting of Uttaranchal , Sikkim and Arunachal Pradesh out of the existing Montane climate region. This we attribute it to the climate change and further investigations are required to ascertain these findings.

The cluster centroids for the seven regions are shown in Table 1. The air temperature in the montanenew regions is very high when compared to Montane region the reasons for under investigation.

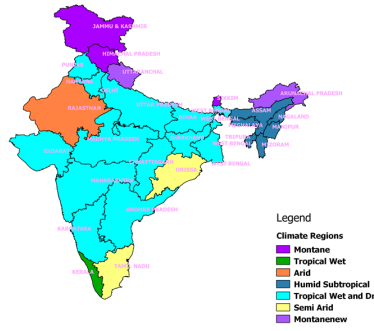


Figure 1: Climate Regions Obtained using Expectation Maximization Clustering Procedure © Dr. M. Naresh Kumar, 2013

Figure 2: Climate Regions Obtained using Expectation Maximization Clustering Procedure

The spatial extents of the climate regions obtained from proposed algorithm 1 is shown in Figure 2.

Table 2: RMSE error for different climate zones		
Region	EM+SVM	KM+LR
Montanenew	1.19	2.9
Semi Arid	0.97	0.88
Tropical Wet and Dry	3.42	2.94
Arid	0.68	0.74
Montane	1.93	1.69
Tropical Wet	0.55	0.48
Humid Sub Tropical	0.8	0.95

The RMSE errors in predicting the temperature for the year 2012 is given in Table 2.

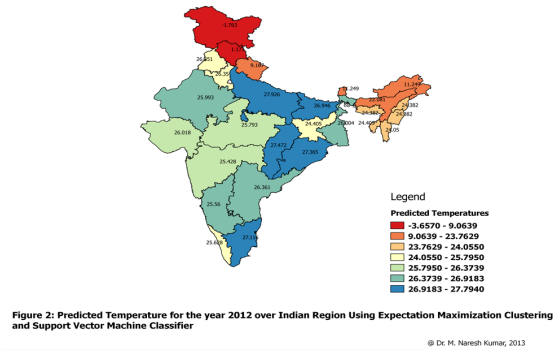


Figure 3: Predicted temperature for the year 2012 over Indian region obtained using Expectation Maximization and SVM Regression Procedure

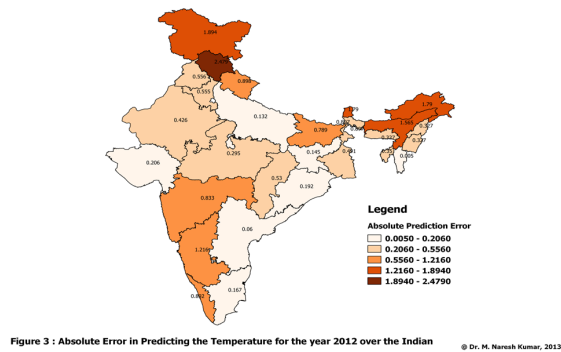


Figure 4: Absolute error in for predicting the temperature the year 2012 over Indian region

The spatial maps of the predicted temperature and the absolute error for the year 2012 over Indian region is shown in Figures 3,4.

The EM clustering could identify all the zones as per the Koppen classification over Indian region. The proposed strategy when employed for predicting temperature has resulted in an RMSE of 1.19 in the Montane climate region and 0.89 in the Humid Sub Tropical region as compared to 2.9 and 0.95 respectively predicted using k-means and linear regression method.

#### **4. Conclusions and Discussion**

The expectation maximization clustering could identify the different climate zones as per the Koppen classification over Indian region. It is observed that the regions of Uttaranchal , Sikkim and Arunachal Pradesh have been identified as a separate group by EM different from the Montane climate zone as per Koppen classification. This needs further investigations and introspection.

EM clustering and SVM performed better than k-means and linear regressions only in Humid subtropical and Montane climate zones. It is observed the EM performance degrades as the dimensionality of the data set increases due to numerical precision problems.

The fast growing volume of climate datasets and its high-dimensionality requires development of novel methods for preprocessing and information extraction. The focus of our future work would be on the development of techniques for big data climate data analytics.

## References

- [1] Chih-Chung Chang, Chih-Jen Lin (2001). LIBSVM - A Library for Support Vector Machines. URL <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [2] Shailesh Kumar Kharol, M. Naresh Kumar, Anu Rani Sharma, Biswadip Gharai, K.V.S. Badarinath, M. Y. Aslam and M. Sivaprasad Reddy. Aerosol Radiative Forcing Over Indian Region- A Study Using CERES, MISR and MODIS Data. In: Proceedings of Aerosols & Clouds : Climate Change Perspectives, 2010. Bose Institute: Darjeeling Campus
- [3] M. Naresh Kumar, Murthy, C. S., Sesha Sai, M. V. R. and Roy, P. S. (2012), Spatiotemporal analysis of meteorological drought variability in the Indian region using standardized precipitation index. Met. Apps vol. 19, pp.256264.
- [4] Mingzhong Xiao, Qiang Zhang, Vijay P. Singh, Xiaohong Chen (2013) Regionalization-based spatiotemporal variations of precipitation regimes across China, Theoretical and Applied Climatology vol. 114, no. 1-2, pp 203-212.
- [5] Srinivas V V, Regionalization of precipitation in India A Review (2013), Journal of the Indian Institute of Science vol 93, no 2.
- [6] Xianliang Zhang, Xiaodong Yan (2013) Temporal change of climate zones in China in the context of climate warming, Theoretical and Applied Climatology
- [7] Zhenxin Bao, Jianyun Zhang, Jiufu Liu, Guobin Fu, Guoqing Wang, Ruimin He, Xiaolin Yan, Junliang Jin, Hongwei Liu (2012) Comparison of regionalization approaches based on regression and similarity for predictions in ungauged catchments under multiple hydro-climatic conditions, Journal of Hydrology vol. 466467, pp 37-46.